A Deep Neural Network Improves Fracture Detection by Clinicians

***Michael J. Gardner, MD*[1]*; Christopher Searles Smith; Timothy S. Achor;*
David Stephenson Wellman, MD; Robert V. O'Toole, MD; Robert N. Hotchkiss;
Aaron Daluiski, MD; Thomas Hotchkiss; Robert Lindsey
[1]*Stanford University, Redwood City, California, USA*

Purpose: Some clinicians lack the subspecialized expertise necessary to identify fractures on radiographs, resulting in missed fracture detection rates as high as 15%. Recent advances in deep learning have produced computer models that learn by example and are effective at many visual identification tasks. Because the models learn by example, subspecialized experts can in principle teach them to detect pathologies by labeling large datasets of radiographs. We hypothesized that (1) a trained algorithm's diagnostic accuracy may be comparable (area under the curve [AUC] ≥0.90) to that of experienced orthopaedic surgeons and (2) a deep learning model taught to detect fractures would improve the diagnostic accuracy of less experienced clinicians.

Methods: To create training examples for the model, 18 senior subspecialized orthopaedic surgeons identified and localized fractures in 135,409 radiographs. We developed and trained a deep neural network model on the radiographs, and we tested its ability to identify fractures on two datasets of wrist radiographs from an academic hospital. We then tested 21 urgent care attending physicians and physician assistants using a within-subjects design to evaluate their diagnostic accuracy with and without the assistance of the model.

Results: On the 2 test sets used for model evaluation, the model's diagnostic accuracy as measured by the area under the receiver operating characteristic curve (AUC ROC) was 0.97 and 0.98. With the assistance of the model, the average urgent care clinician's sensitivity increased from 79.0% to 90.1% (2-sided Wilcoxon signed rank test, P <0.0001, Cohen's d = 1:14) and specificity increased from 85.9% to 94.5% (P <0.0001; d = 1:42). The average clinician experienced a relative reduction in misinterpretation rate of 52.0% (95% confidence interval [CI], 38.0% - 60.8%). The model achieved 93.9% sensitivity (95% CI, 82.9% - 98.0%), 94.5% specificity (95% CI, 90.9% - 96.8%), and .990 AUC (95% CI, .975 - .996) on the same radiographs.

Conclusion: We have demonstrated that a machine learning algorithm trained on a large dataset can produce a fracture detector that outputs heat maps and detects fractures with diagnostic accuracy comparable to experienced orthopaedic surgeons. Because the software takes milliseconds to make an assessment and can run on most computers, it has the potential to significantly reduce the incidence of missed fractures and improve patient care worldwide. When the model's output was provided to urgent care clinicians, their diagnostic accuracy was significantly improved.